Contents

Copyrighted Material

Preface

> The greatest possibilities of visual display lie in vividness and inescapability of the intended message. A visual display can stop your mental flow in its tracks and make you think. A visual display can force you to notice what you never expected to see.
>
> John W. Tukey (1990)

Data analysis and graphics

This book stems from the conviction that data analysis and statistical graphics should go hand-in-hand in the process of understanding and communicating statistical data. Statistical summaries compress a data set into a few numbers, the result of an hypothesis test, or coefficients in a fitted statistical model, while graphical methods help us to explore patterns and trends, see the unexpected, identify problems in an analysis, and communicate results and conclusions in principled and effective ways.

This interplay between analysis and visualization has long been a part of statistical practice for *quantitative data*. Indeed, the origin of correlation, regression, and linear models (regression, ANOVA) can arguably be traced to Francis Galton's (1886) visual insight from a scatterplot of heights of children and their parents on which he overlaid smoothed contour curves of roughly equal bivariate frequencies and lines for the means of $Y \mid X$ and $X \mid Y$ (described in Friendly and Denis (2005), Friendly et al. (2013)).

The analysis of discrete data is a much more recent arrival, beginning in the 1960s and giving rise to a few seminal books in the 1970s (Bishop et al., 1975, Haberman, 1974, Goodman, 1978, Fienberg, 1980). Agresti (2013, Chapter 17) presents a brief historical overview of the development of these methods from their early roots around the beginning of the 20^{th} century.

Yet curiously, associated graphical methods for categorical data were much slower to develop. This began to change as it was recognized that counts, frequencies, and discrete variables required different schemes for mapping numbers into useful visual representations (Friendly, 1995, 1997), some quite novel. The special nature of discrete variables and frequency data vis-a-vis statistical

graphics is now more widely accepted, and many of these new graphical methods (e.g., mosaic displays, fourfold plots, diagnostic plots for generalized linear models) have become, if not mainstream, then at least more widely used in research, teaching, and communication.

Much of what had been developed through the 1990s for graphical methods for discrete data was

described in the book *Visualizing Categorical Data* (Friendly, 2000) and was implemented in SAS® software. Since that time, there has been considerable growth in both statistical methods for the analysis of categorical data (e.g., generalized linear models, zero-inflation models, mixed models for hierarchical and longitudinal data with discrete outcomes), along with some new graphical methods for visualizing and interpreting the results (3D mosaic plots, effect plots, diagnostic plots, etc.). The bulk of these developments have been implemented in R, and the time is right for an in-depth treatment of modern graphical methods for the analysis of categorical data, to which you are now invited.

Goals

This book aims to provide an applied, practically oriented treatment of modern methods for the analysis of categorical data—discrete response data and frequency data—with a main focus on graphical methods for exploring data, spotting unusual features, visualizing fitted models, and presenting or explaining results.

We describe the necessary statistical theory (sometimes in abbreviated form) and illustrate the practical application of these techniques to a large number of substantive problems: how to organize the data, conduct an analysis, produce informative graphs, and understand what they have to say about the data at hand.

Overview and organization of this book

This book is divided into three parts. Part I, Chapters 1–3, contains introductory material on graphical methods for discrete data, basic R skills needed for the book, and methods for fitting and visualizing one-way discrete distributions.

Part II, Chapters 4–6, is concerned largely with simple, traditional non-parametric tests and exploratory methods for visualizing patterns of association in two-way and larger frequency tables. Some of the discussion here introduces ideas and notation for loglinear models that are treated more generally in Part III.

Part III, Chapters 7–11, discusses model-based methods for the analysis of discrete data. These are all examples of generalized linear models. However, for our purposes, it has proved more convenient to develop this topic from the specific cases (logistic regression, loglinear models) to the general rather than the reverse.

Chapter 1: ***Introduction.*** Categorical data require different statistical and graphical methods than commonly used for quantitative data. This chapter outlines the basic orientation of the book toward visualization methods and some key distinctions regarding the analysis and visualization of categorical data.

Chapter 2: ***Working with Categorical Data.*** Categorical data can be represented in various forms: case form, frequency form, and table form. This chapter describes and illustrates the skills and techniques in R needed to input, create, and manipulate R data objects to represent categorical data, and convert these from one form to another for the purposes of statistical analysis and visualization, which are the subject of the remainder of the book.

Chapter 3: ***Fitting and Graphing Discrete Distributions.*** Understanding and visualizing discrete data distributions provides a building block for model-based methods discussed in Part III. This chapter introduces the well-known discrete distributions—the binomial, Poisson, negative-binomial, and others—in the simplest case of a one-way frequency table.

Chapter 4: *Two-Way Contingency Tables*. The analysis of two-way frequency tables concerns the association between two variables. A variety of specialized graphical displays help to visualize the pattern of association, using area of some region to represent the frequency in a cell. Some of these methods are focused on visualizing an odds ratio (for 2×2 tables), or the general pattern of association, or the agreement between row and column categories in square tables.

Chapter 5: *Mosaic Displays for n-Way Tables*. This chapter introduces mosaic displays, designed to help to visualize the pattern of associations among variables in two-way and larger tables. Extensions of this technique can reveal partial associations and marginal associations, and shed light on the structure of loglinear models themselves.

Chapter 6: *Correspondence Analysis*. Correspondence analysis provides visualizations of associations in a two-way contingency table in a small number of dimensions. Multiple correspondence analysis extends this technique to n-way tables. Other graphical methods, including mosaic matrices and biplots, provide complementary views of loglinear models for two-way and n-way contingency tables.

Chapter 7: *Logistic Regression Models*. This chapter introduces the modeling framework for categorical data in the simple situation where we have a categorical response variable, often binary, and one or more explanatory variables. A fitted model provides both statistical inference and prediction, accompanied by measures of uncertainty. Data visualization methods for discrete response data must often rely on smoothing techniques, including both direct, non-parametric smoothing and the implicit smoothing that results from a fitted parametric model. Diagnostic plots help us to detect influential observations that may distort our results.

Chapter 8: *Models for Polytomous Responses*. This chapter generalizes logistic regression models for a binary response to handle a multi-category (polytomous) response. Different models are available depending on whether the response categories are nominal or ordinal. Visualization methods for such models are mostly straightforward extensions of those used for binary responses presented in Chapter 7.

Chapter 9: *Loglinear and Logit Models for Contingency Tables*. This chapter extends the model-building approach to loglinear and logit models. These comprise another special case of generalized linear models designed for contingency tables of frequencies. They are most easily interpreted through visualizations, including mosaic displays and effect plots of associated logit models.

Chapter 10: *Extending Loglinear Models*. Loglinear models have special forms to represent additional structure in the variables in contingency tables. Models for ordinal factors allow a more parsimonious description of associations. Models for square tables allow a wide range of specific models for the relationship between variables with the same categories. Another extended class of models arise when there are two or more response variables.

Chapter 11: *Generalized Linear Models*. Generalized linear models extend the familiar linear models of regression and ANOVA to include counted data, frequencies, and other data for which the assumptions of independent, normal errors are not reasonable. We rely on the analogies between ordinary and generalized linear models (GLMs) to develop visualization methods to explore the data, display the fitted relationships, and check model assumptions. The main focus of this chapter is on models for count data.

Audience

This book has been written to appeal to two broad audiences wishing to learn to apply methods for discrete data analysis:

- Advanced undergraduate and graduate students in the social and health sciences, epidemiology, economics, business, and (bio)statistics

- Substantive researchers, methodologists, and consultants in various disciplines wanting to be able to use these methods with their own data and analyses.

It assumes the reader has a basic understanding of statistical concepts at least at an intermediate undergraduate level including regression and analysis of variance (for example, at the level of Neter et al. (1990) or Mendenhall and Sincich (2003)). It is less technically demanding than other modern texts covering categorical data analysis at a graduate level, such as Agresti (2013), *Categorical Data Analysis*, Powers and Xie (2008), *Statistical Methods for Categorical Data Analysis*, and Christensen (1997), *Log-Linear Models and Logistic Regression*. Nevertheless, there are some topics that are a bit more advanced or technical, and these are marked as * or ** sections.

As well, there are a number of mathematical or statistical topics that we use in passing, but do not describe in these pages (some matrix notation, basic probability theory, maximum likelihood estimation, etc.). Most of these are described in Fox (2015), which is available online and serves well as a supplement to this book.

In addition, it is not possible to include *all* details of using R effectively for data analysis. It is assumed that the reader has at least basic knowledge of the R language and environment, including interacting with the R console (RGui for Windows, R.app for Mac OS X) or other graphical user interface (e.g., RStudio), using R functions in packages, getting help for these from R, etc. One introductory chapter (Chapter 2) is devoted to covering the particular topics most important to categorical data analysis, beyond such basic skills needed in the book.

Textbook use

This book is most directly suitable for a one-semester applied advanced undergraduate or graduate course on categorical data analysis with a strong emphasis on the use of graphical methods to understand and explain data and results of analysis. A detailed outline of such a course, together with lecture notes and assignments, is available at the first author's web page, `http://euclid.psych.yorku.ca/www/psy6136/`, using this book as the main text. This course also uses Agresti (2007), *An Introduction to Categorical Data Analysis* for additional readings.

For instructors teaching a more traditional course using one of the books mentioned above as the main text, this book would be a welcome supplement, because almost all other texts treat graphical methods only perfunctorily, if at all. A few of these contain a brief appendix mentioning software, or have a related web site with some data sets and software examples. Moreover, none actually describe how to do these analyses and graphics with R.

Features

- Provides an accessible introduction to the major methods of categorical data analysis for data exploration, statistical testing, and statistical models.
- The emphasis throughout is on computing, visualizing, understanding, and communicating the results of these analyses.
- As opposed to more theoretical books, the goal here is to help the reader to translate theory into practical application, by providing skills and software tools for carrying out these methods.
- Includes many examples using real data, often treated from several perspectives.
- The book is supported directly by R packages vcd (Meyer et al., 2015) and vcdExtra (Friendly, 2015), along with numerous other R packages.
- All materials (data sets, R code) will be available online on the web site for the book, `http://datavis.ca/books/DDAR`.

Copyrighted Material

- Each chapter contains a collection of lab exercises, which work through applications of some of the methods presented in that chapter. This makes the book more suitable for both self-study and classroom use.

Acknowledgments

We are grateful to many colleagues, friends, students, and Internet acquaintences who have contributed to to this book, directly or indirectly.

We thank those who read and commented on various drafts of the book or chapters. In particular, John Fox, Michael Greenacre, and several anonymous reviewers gave insightful comments on the

organization of the book and made many helpful suggestions. Matthew Sigal used his wizardly skills to turn sketches of conceptual diagrams into final figures. Phil Chalmers contributed greatly with technical and stylistic editing of a number of chapters.

At a technical level, we were aided by the cooperation of a number of R package authors, who helped to enhance the graphic displays: Achim Zeilleis who served as a guiding hand in the development of the vcd and vcdExtra packages; John Fox and Sandy Weisberg for enhancements to the car (Fox and Weisberg, 2015a) and effects (Fox et al., 2015) packages; Milan Bouchet-Valat for incorporating suggestions dealing with plotting `rc()` solutions into the logmult (Bouchet-Valat, 2015) package; Michael Greenacre and Oleg Nenadic for help to enhance plotting in the ca (Greenacre and Nenadic, 2014) package; Heather Turner for advice and help with plotting models fit using the gnm (Turner and Firth, 2014) package; Jay Emerson for improvements to the gpairs (Emerson and Green, 2014) package.

There were also many contributors from the R-Help email list (`r-help@r-project.org`), too many to name them all. Special thanks for generous assistance go to: David Carlson, William Dunlap, Bert Gunter, Jim Lemon, Duncan Murdoch, Denis Murphy, Jeff Newmiller, Richard Heiberger, Thierry Onkelinx, Marc Schwartz, David Winsemius, and Ista Zahn.

The book was written using the knitr (Xie, 2015) package, allowing a relatively seamless in-

1

1.3 Strategies for analysis

Introduction

Categorical data consist of variables whose values comprise a set of discrete categories. Such data require different statistical and graphical methods than commonly used for quantitative data. The focus of this book is on visualization techniques and graphical methods designed to reveal patterns of relationships among categorical variables. This chapter outlines the basic orientation of the book and some key distinctions regarding the analysis and visualization of categorical data.

1.1 Data visualization and categorical data: Overview

> Graphs carry the message home. A universal language, graphs convey information directly to the mind. Without complexity there is imaged to the eye a magnitude to be remembered. Words have wings, but graphs interpret. Graphs are pure quantity, stripped of verbal sham, reduced to dimension, vivid, unescapable.
>
> Henry D. Hubbard, in Foreword to Brinton (1939), *Graphic Presentation*

"Data visualization" can mean many things, from popular press infographics, to maps of voter turnout or party choice. Here we use this term in the narrower context of statistical analysis. As such, we refer to an approach to data analysis that focuses on *insightful* graphical display in the service of both *understanding* our data and *communicating* our results to others.

We may display the raw data, some summary statistics, or some indicators of the quality or adequacy of a fitted model. The word "insightful" suggests that the goal is (hopefully) to reveal some aspects of the data that might not be perceived, appreciated, or absorbed by other means. As

in the quote from Keats, the overall aims include both beauty and truth, though each of these are

only as perceived by the beholder.

Methods for visualizing quantitative data have a long history and are now widely used in both data analysis and in data presentation, and in both popular and scientific media. Graphical methods for categorical data, however, have only a more recent history, and are consequently not as widely used. The goal of this book is to show concretely how data visualization may be usefully applied to categorical data.

"Categorical" means different things in different contexts. We introduce the topic in Section 1.2 with some examples illustrating (a) types of categorical variables: binary, nominal, and ordinal, (b) data in case form vs. frequency form, (c) frequency data vs. count data, (d) univariate, bivariate, and multivariate data, and (e) the distinction between explanatory and response variables.

Statistical methods for the analysis of categorical data also fall into two quite different categories, described and illustrated in Section 1.3: (a) the simple randomization-based methods typified by the classical Pearson chi-squared (χ^2) test, Fisher's exact test, and Cochran–Mantel–Haenszel tests, and (b) the model-based methods represented by logistic regression, loglinear, and generalized linear models. In this book, Chapters 3–6 are mostly related to the randomization-based methods; Chapters 7–9 illustrate the model-based methods.

In Section 1.4 we describe some important similarities and differences between categorical data and quantitative data, and discuss the implications of these differences for visualization techniques. Section 1.4.5 outlines a strategy of data analysis focused on visualization.

In a few cases we show R code or results as illustrations here, but the fuller discussion of using R for categorical data analysis is postponed to Chapter 2.

1.2 What is categorical data?

A ***categorical variable*** is one for which the possible measured or assigned values consist of a discrete set of categories, which may be *ordered* or *unordered*. Some typical examples are:

- `Gender`, with categories "Male," "Female."
- `Marital status`, with categories "Never married," "Married," "Separated," "Divorced," "Widowed."
- `Fielding position` (in baseball), with categories "Pitcher," "Catcher," "1st base," "2nd base," ..., "Left field."
- `Side effects` (in a pharmacological study), with categories "None," "Skin rash," "Sleep disorder," "Anxiety," ....
- `Political attitude`, with categories "Left," "Center," "Right."
- `Party preference` (in Canada), with categories "NDP," "Liberal," "Conservative," "Green."
- `Treatment outcome`, with categories "no improvement," "some improvement," or "marked improvement."
- `Age`, with categories "0–9," "10–19," "20–29," "30–39," ....
- `Number of children`, with categories $0, 1, 2, \ldots$.

As these examples suggest, categorical variables differ in the number of categories: we often distinguish ***binary variables*** (or ***dichotomous variables***) such as `Gender` from those with more than two categories (called ***polytomous variables***). For example, Table 1.1 gives data on $4,526$ applicants to graduate departments at the University of California at Berkeley in 1973, classified by two binary variables, gender and admission status.

Some categorical variables (`Political attitude`, `Treatment outcome`) may have ordered categories (and are called ***ordinal variables***), while others (***nominal variables***) like `Marital`

Table 1.1: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total
Males	1198	1493	2691
Females	557	1278	1835
Total	1755	2771	4526

Table 1.2: Arthritis treatment data

		Improvement			
Treatment	Sex	None	Some	Marked	Total
Active	Female	6	5	16	27
	Male	7	2	5	14
Placebo	Female	19	7	6	32
	Male	10	0	1	11
Total		42	14	28	84

`status` have unordered categories.[1] For example, Table 1.2 shows a $2 \times 2 \times 3$ table of ordered outcomes ("none," "some," or "marked" improvement) to an active treatment for rheumatoid arthritis compared to a placebo for men and women.

Finally, such variables differ in the fineness or level to which some underlying observation has been categorized for a particular purpose. From one point of view, *all* data may be considered categorical because the precision of measurement is necessarily finite, or an inherently continuous variable may be recorded only to limited precision.

But this view is not helpful for the applied researcher because it neglects the phrase "for a particular purpose." Age, for example, might be treated as a quantitative variable in a study of native language vocabulary, or as an ordered categorical variable with decade groups (0–10, 11–20, 20–30, ...) in terms of the efficacy or side-effects of treatment for depression, or even as a binary variable ("child" vs. "adult") in an analysis of survival following an epidemic or natural disaster. In the analysis of data using categorical methods, continuous variables are often recoded into ordered categories with a small set of categories for some purpose.[2]

1.2.1 Case form vs. frequency form

In many circumstances, data is recorded on each individual or experimental unit. Data in this form is called case data, or data in ***case form***. The data in Table 1.2, for example, were derived from the individual data listed in the data set `Arthritis` from the vcd package. The following lines show the first five of $N = 84$ cases in the `Arthritis` data,

```
  ID Treatment  Sex Age Improved
1 57   Treated Male  27     Some
```

[1] An ordinal variable may be defined as one whose categories are *unambiguously* ordered along a *single* underlying dimension. Both marital status and fielding position may be weakly ordered, but not on a single dimension, and not unambiguously.

[2] This may be a waste of information available in the original variable, and should be done for substantive reasons, not mere convenience. For example, some researchers unfamiliar with regression methods often perform a "median-split" on quantitative predictors so they can use ANOVA methods. Doing this precludes the possibility of determining if those variables have nonlinear relations with the outcome while also decreasing statistical power.


```
2 46   Treated Male  29     None
3 77   Treated Male  30     None
4 17   Treated Male  32   Marked
5 36   Treated Male  46   Marked
```

Whether or not the data variables, and the questions we ask, call for categorical or quantitative data analysis, when the data are in case form, we can always trace any observation back to its individual identifier or data record (for example, if the case with `ID` equal to 57 turns out to be unusual or noteworthy).

Data in ***frequency form*** has already been tabulated, by counting over the categories of the table variables. The same data shown as a table in Table 1.2 appear in frequency form as shown below.

```
   Treatment    Sex Improved Freq
1    Placebo Female     None   19
2    Treated Female     None    6
3    Placebo   Male     None   10
4    Treated   Male     None    7
5    Placebo Female     Some    7
6    Treated Female     Some    5
```

```
7    Placebo   Male     Some    0
8    Treated   Male     Some    2
9    Placebo Female   Marked    6
10   Treated Female   Marked   16
11   Placebo   Male   Marked    1
12   Treated   Male   Marked    5
```

Data in frequency form may be analyzed by methods for quantitative data if there is a quantitative response variable (weighting each group by the cell frequency, with a weight variable). Otherwise, such data are generally best analyzed by methods for categorical data, where statistical models are often expressed as models for the frequency variable, in the form of an R formula like `Freq ~ ..`

In any case, an observation in a data set in frequency form refers to all cases in the cell collectively, and these cannot be identified individually. Data in case form can always be reduced to frequency form, but the reverse is rarely possible. In Chapter 2, we identify a third format, ***table form***, which is the R representation of a table like Table 1.2.

1.2.2 Frequency data vs. count data

In many cases the observations representing the classifications of events (or variables) are recorded from *operationally independent* experimental units or individuals, typically a sample from some population. The tabulated data may be called ***frequency data***. The data in Table 1.1 and Table 1.2 are both examples of frequency data because each tabulated observation comes from a different person.

However, if several events or variables are observed for the same units or individuals, those events are not operationally independent, and it is useful to use the term ***count data*** in this situation. These terms (following Lindsey (1995)) are by no means standard, but the distinction is often important, particularly in statistical models for categorical data.

For example, in a tabulation of the number of male children within families (Table 1.3, described in Section 1.2.3 below), the number of male children in a given family would be a *count* variable, taking values $0, 1, 2, \ldots$. The number of independent families with a given number of male children is a *frequency* variable. Count data also arise when we tabulate a sequence of events over time or under different circumstances in a number of individuals.

Table 1.3: Number of Males in 6115 Saxony Families of Size 12

Males	0	1	2	3	4	5	6	7	8	9	10	11	12
Families	3	24	104	286	670	1,033	1,343	1,112	829	478	181	45	7

1.2.3 Univariate, bivariate, and multivariate data

Another distinction concerns the number of variables: one, two, or (potentially) many shown in a data set or table, or used in some analysis. Table 1.1 is an example of a bivariate (two-way) contingency table and Table 1.2 classifies the observations by three variables. Yet, we will see later that the Berkeley admissions data also recorded the department to which potential students applied (giving a three-way table), and in the arthritis data, the age of subjects was also recorded.

Any contingency table (in frequency or table form) therefore records the *marginal totals*, summed over all variables not represented in the table. For data in case form, this means simply ignoring (or not recording) one or more variables; the "observations" remain the same. Data in frequency form, however, result in smaller tables when any variable is ignored; the "observations" are the cells of the contingency table. For example, in the `Arthritis` data, ignoring `Sex` gives the smaller 2×3 table for `Treatment` and `Improved`.

```
  Treatment Improved Freq
1   Placebo     None   29
2   Treated     None   13
3   Placebo     Some    7
4   Treated     Some    7
```

```
5   Placebo   Marked    7
6   Treated   Marked   21
```

In the limiting case, only one table variable may be recorded or available, giving the categorical equivalent of univariate data. For example, Table 1.3 gives data on the distribution of the number of male children in families with 12 children (discussed further in Example 3.2). These data were part of a large tabulation of the sex distribution of families in Saxony in the 19^{th} century, but the data in Table 1.3 have only one discrete classification variable, number of males. Without further information, the only statistical questions concern the form of the distribution. We discuss methods for fitting and graphing such discrete distributions in Chapter 3. The remaining chapters relate to bivariate and multivariate data.

1.2.4 Explanatory vs. response variables

Most statistical models make a distinction between ***response variables*** (or *dependent*, or *criterion* variables) and ***explanatory variables*** (or *independent*, or *predictor* variables).

In the standard (classical) linear models for regression and analysis of variance (ANOVA), for instance, we treat one (or more) variables as responses, to be explained by the other, explanatory variables. The explanatory variables may be quantitative or categorical (e.g., factors in R). This affects only the details of how the model is specified or how coefficients are interpreted for `lm()` or `glm()`. In these classical models, the response variable ("treatment outcome," for example), must be considered quantitative, and the model attempts to describe how the *mean* of the distribution of responses changes with the values or levels of the explanatory variables, such as age or gender.

When the response variable is categorical, however, the standard linear models do not apply, because they assume a normal (Gaussian) distribution for the model residuals. For example, in Table 1.2 the response variable is `Improvement`, and even if numerical scores were assigned to

the categories "none," "some," "marked," it may be unlikely that the assumptions of the classical linear models could be met.

Hence, a categorical *response* variable generally requires analysis using methods for categorical data, but categorical *explanatory* variables may be readily handled by either method.

The distinction between response and explanatory variables also becomes important in the use of loglinear models for frequency tables (described in Chapter 9), where models can be specified in a simpler way (as equivalent logit models) by focusing on the response variable.

1.3 Strategies for categorical data analysis

Data analysis typically begins with exploratory and graphical methods designed to expose features of the data, followed by statistical analysis designed to summarize results, answer questions, and draw conclusions. Statistical methods for the analysis of categorical data can be classified into two broad categories: those concerned with *hypothesis testing* per se versus those concerned with *model building*.

1.3.1 Hypothesis testing approaches

In many studies, the questions of substantive interest translate readily into questions concerning hypotheses about ***association*** between variables, a more general idea than that of correlation (*linear* association) for quantitative variables. If a non-zero association exists, we may wish to characterize the strength of the association numerically and understand the pattern or nature of the association.

For example, in Table 1.1, a main question is: "Is there evidence of gender-bias in admission to graduate school?" Another way to frame this: "Are males more likely to be admitted?" These questions can be expressed in terms of an association between gender and admission status in a 2×2 contingency table of applicants classified by these two variables. If there is evidence for an association, we can assess its strength by a variety of measures, including the difference in proportions admitted for men and women or the ratio of the odds of admission for men compared to women, as described in Section 4.2.2.

Similarly, in Table 1.2, questions about the efficacy of the treatment for rheumatoid arthritis can be answered in terms of hypotheses about the associations among the table variables: `Treatment`, `Sex` and the `Improvement` categories. Although the main concern might be focused on the

overall association between `Treatment` and `Improvement`, one would also wish to know if this association is the same for men and women. A ***stratified analysis*** (Section 4.3) controls for the effects of background variables like Sex, tests for ***homogeneity of association***, and helps to determine if these associations are equal.

Questions involving tests of such hypotheses are answered most easily using a large variety of specific statistical tests, often based on randomization arguments. These include the familiar Pearson chi-squared test for two-way tables, the Cochran–Mantel–Haenszel test statistics, Fisher's exact test, and a wide range of measures of strength of association. These tests make minimal assumptions, principally requiring that subjects or experimental units have been randomly assigned to the categories of experimental factors. The hypothesis testing approach is illustrated in Chapters 4–6, though the emphasis is on graphical methods that help us to understand the nature of association between variables.

EXAMPLE 1.1: Hair color and eye color

The data set `HairEye` below records data on the relationship between hair color and eye color in a sample of nearly 600 students.

Copyrighted Material

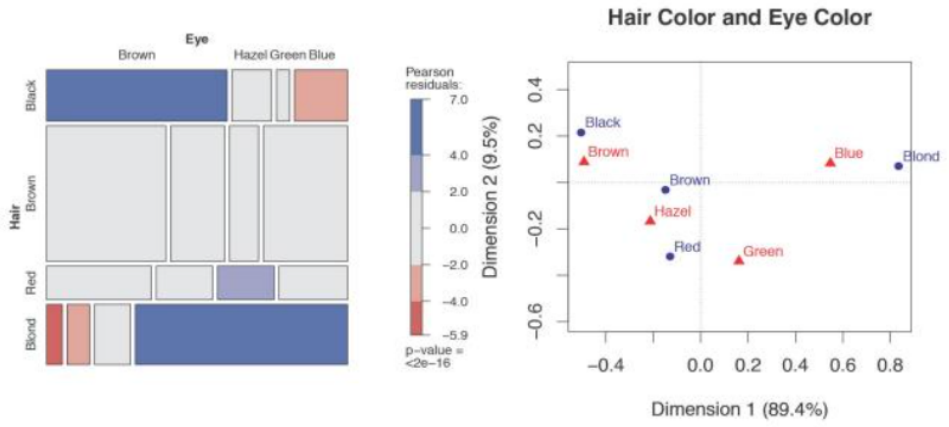

Figure 1.1: Graphical displays for the hair color and eye color data. Left: mosaic display; right: correspondence analysis plot.

```
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16
```

The standard analysis (with `chisq.test()` or `assocstats()`) gives a Pearson χ^2 of 138.3 with nine degrees of freedom, indicating substantial departure from independence. Among the measures of strength of association, ***Cramer's V***, $V = \sqrt{\chi^2/N \min(r-1, c-1)} = 0.279$, indicates a substantial relationship between hair and eye color.[3]

```
                    X^2 df P(> X^2)
Likelihood Ratio 146.44  9        0
Pearson          138.29  9        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.435
Cramer's V        : 0.279
```

The further (and perhaps more interesting question) is how do we understand the *nature* of this association between hair and eye color? Two graphical methods related to the hypothesis testing approach are shown in Figure 1.1.

The left panel of Figure 1.1 is a ***mosaic display*** (Chapter 5), constructed so that the size of each rectangle is proportional to the observed cell frequency. The shading reflects the cell contribution to the χ^2 statistic—shades of blue when the observed frequency is substantially greater than the expected frequency under independence, shades of red when the observed frequency is substantially less, as shown in the legend.

The right panel of this figure shows the results of a correspondence analysis (Chapter 6), where the deviations of the hair color and eye color points from the origin accounts for as much of the χ^2 as possible in two dimensions.

[3]Cramer's V varies from 0 (no association) to 1 (perfect association).

We observe that both the hair colors and the eye colors are ordered from dark to light in the mosaic display and along Dimension 1 in the correspondence analysis plot. The deviations between observed and expected frequencies have an opposite-corner pattern in the mosaic display, except for the combination of red hair and green eyes, which also stand out as the largest values on Dimension 2 in the Correspondence analysis plot. Displays such as these provide a means to understand *how* the variables are related. △

1.3.2 Model building approaches

Model-based methods provide tests of equivalent hypotheses about associations, but offer additional advantages (at the cost of additional assumptions) not provided by the simpler hypotheses-testing approaches. Among these advantages, model-based methods provide estimates, standard errors and confidence intervals for parameters, and the ability to obtain predicted (fitted/expected) values with associated measures of precision.

We illustrate this approach here for a dichotomous response variable, where it is often convenient to construct a model relating a function of the probability, π, of one event to a linear combination of the explanatory variables. Logistic regression uses the ***logit function***,

$$\text{logit}(\pi) \equiv \log_e \left(\frac{\pi}{1-\pi} \right) ,$$

which may be interpreted as the ***log odds*** of the given event. A linear logistic model can then be expressed as

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

Statistical inferences from model-based methods provide tests of hypotheses for the effects of the predictors, $x_1, x_2, \ldots$, but they also provide estimates of parameters in the model, $\beta_1, \beta_2, \ldots$ and associated confidence intervals. Standard modeling tools allow us to graphically display the fitted response surface (with confidence or prediction intervals) and even to extrapolate these predictions beyond the given data. A particular advantage of the logit representation in the logistic regression model is that estimates of odds ratios (Section 4.2.2) may be obtained directly from the parameter estimates.

EXAMPLE 1.2: Space shuttle disaster

To illustrate the model-based approach, the graph in Figure 1.2 is based on a logistic regression model predicting the probability of a failure in one of the O-ring seals used in the 24 NASA space shuttles prior to the disastrous launch of the *Challenger* in January, 1986. The explanatory variable is the ambient temperature (in Fahrenheit) at the time of the flight. The sad story behind these data, and the lessons to be learned for graphical data display, are related in Example 1.10.

Here, we simply note that the fitted model, shown by the solid line in Figure 1.2, corresponds to the prediction equation (with standard errors shown in parentheses),

$$\underset{}{\text{logit}(\text{Failure})} = \underset{(3.06)}{5.09} - \underset{(0.047)}{0.116}\ \text{Temperature}$$

A hypothesis test that failure probability is unassociated with temperature is equivalent to the test that the coefficient for temperature in this model equals 0; this test has a p-value of 0.014, convincing evidence for rejection.

The parameter estimate for temperature, -0.116, however, gives more information. Each 1° increase in temperature decreases the log odds of failure by 0.116, with 95% confidence interval $[-0.208, -0.0235]$. The equivalent odds ratio is $\exp(-0.116) = 0.891$ $[0.812, 0.977]$. Equivalently, a 10° *decrease* in temperature corresponds to an odds ratio of a failure of $\exp(10 \times 0.116) =$

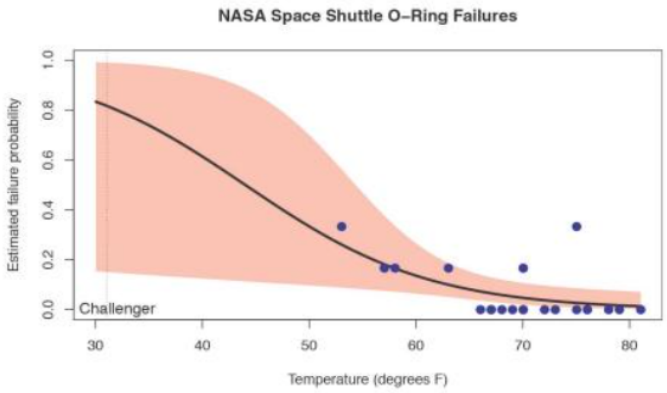

Figure 1.2: Space shuttle O-ring failure, observed and predicted probabilities. The dotted vertical line at 31° shows the prediction for the launch of the *Challenger*.

When the *Challenger* was launched, the temperature was only 31°. The shaded region in Figure 1.2 shows 95% prediction intervals for failure probability. All previous shuttles (shown by the points in the figure) had been launched at much warmer temperatures, so the prediction interval (the dashed vertical line) at 31° represents a considerable extrapolation beyond the available data. Nonetheless, the model building approach does provide such predictions along with measures of their uncertainty. Figure 1.2 is a graph that might have saved lives.

△

EXAMPLE 1.3: Donner Party

In April–May of 1846 (three years before the California gold rush), the Donner and Reed families set out for California from the American Mid-west in a wagon train to seek a new life and perhaps their fortune in the new American frontier. By mid-July, a large group had reached a site in present-day Wyoming; George Donner was elected to lead what was to be called the "Donner Party," which eventually numbered 87 people in 23 wagons, along with their oxen, cattle, horses, and worldly possessions.

They were determined to reach California as quickly as possible. Lansford Hastings, a self-proclaimed trailblazer (retrospectively, of dubious distinction), proposed that the party follow him through a shorter path through the Wasatch Mountains. Their choice of "Hastings's Cutoff" proved disastrous: Hastings had never actually crossed that route himself, and the winter of of 1846 was to be one of the worst on record.

In October, 1846, heavy snow stranded them in the eastern Sierra Nevada, just to the east of a pass that bears their name today. The party made numerous attempts to seek rescue, most turned back by blizzard conditions. Relief parties in March–April 1847 rescued 40, but discovered grisly evidence that those who survived had cannibalized those who died.

Here we briefly examine how statistical models and graphical evidence can shed light on the question of who survived in the Donner party.

Figure 1.3 is an example of what we call a *data-centric, model-based* graph of a discrete (binary) outcome: lived (1) versus died (0). That is, it shows both the data and a statistical summary based on a fitted statistical model. The statistical model provides a smoothing of the discrete data.

The jittered points at the top and bottom of the graph show survival in relation to age of the person. You can see that there were more people who survived among the young, and more who died among the old. The blue curve in the plot shows the fitted probability of survival from a

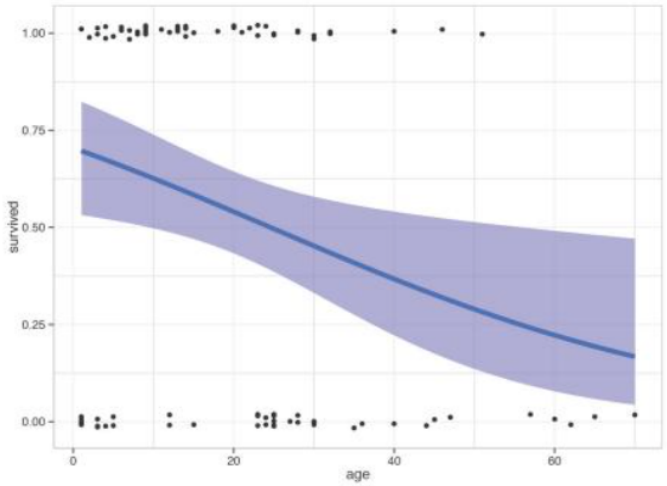

Figure 1.3: Donner party data, showing the relationship between age and survival. The blue curve and confidence band give the predicted probability of survival from a linear logistic regression model.

linear logistic regression model for these data with a 95% confidence band for the predictions. The prediction equation for this model can be given as:

$$\underset{}{\text{logit}(\text{survived})} = \underset{(0.372)}{0.868} - \underset{(0.015)}{0.0353}\,\text{age}$$

The equation above implies that the log odds of survival decreases by 0.0352 with each additional year of age or by $10 \times 0.0352 = 0.352$ for an additional decade. Another way to say this is that the odds of survival is multiplied by $\exp(0.353) = .702$ with each 10 years of age, a 30% decrease.

Of course, these visual and statistical summaries depend on the validity of the fitted model. For

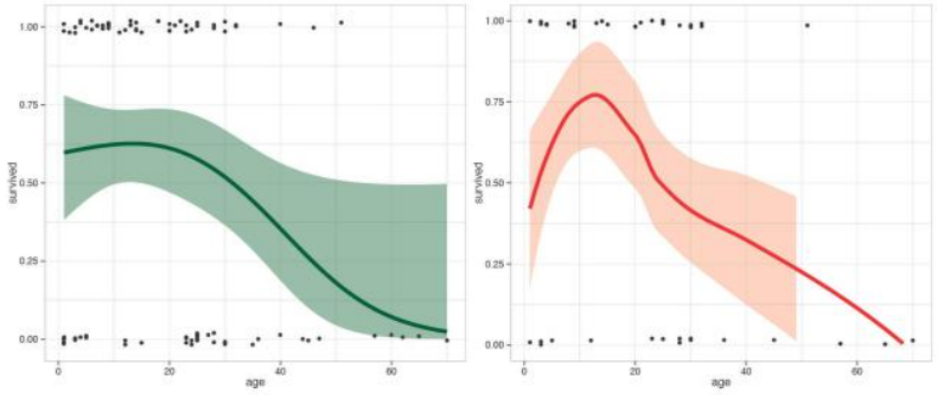

Figure 1.4: Donner party data, showing other model-based smoothers for the relationship between age and survival. Left: using a natural spline; right: using a non-parametric loess smoother.